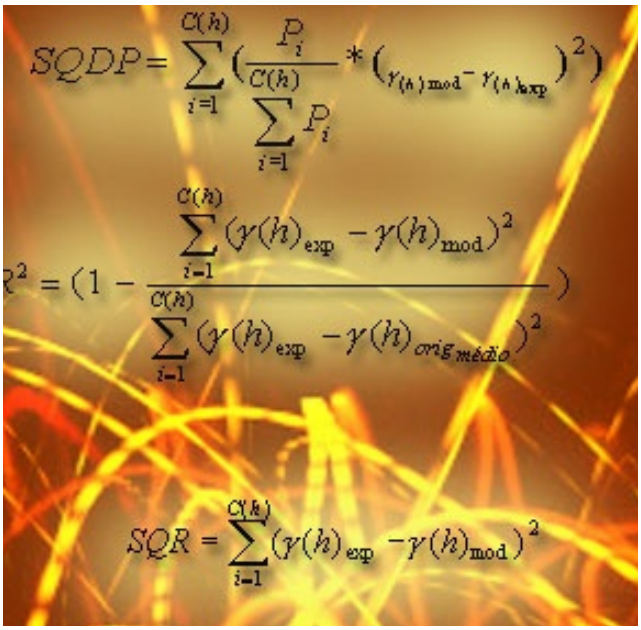
Comunicado Técnico 58
Dezembro, 2003
Campinas, SP



ISSN 1677-8464

Uso de Índices de Desempenho e do Critério de Akaike para Ajuste de Modelos de Semivariograma

Laurimar Gonçalves Vendrusculo[1]

O semivariograma constitui-se numa ferramenta importante para representação quantitativa da variabilidade espacial e temporal de determinado atributo. Segundo Vieira (2000), esta ferramenta é a mais adequada para medir dependência espacial e interpolação em locais não-amostrados com intuito de gerar mapas de isolinhas ou tridimensionais para exame e interpretação de variabilidade.

Em geoestatística a escolha do melhor modelo de semivariograma é crucial para a correta interpretação de fenômenos que apresentem dependência espacial. A esta escolha, repercute diretamente a confiabilidade dos resultados oriundos do processo de interpolação (krigagem); ou seja, na confiabilidade dos valores estimados em pontos não-amostrados.

De uma maneira geral, a modelagem de sistemas tais como os biológicos e químicos, constituem-se em uma atividade que contempla múltiplos e conflitantes objetivos. Estes objetivos envolvem a complexidade do modelos, o(s) critério(s) de desempenho e a validação, que influenciam a seleção da estrutura do modelo matemático. Alguns príncipios podem auxiliar a construção destes modelos, um deles é o princípio da parcimônia onde os melhores modelos são obtidos utilizando-se estruturas aceitáveis e simples, contendo em sua formulação um menor número de parâmetros. Lark (2001) usou este princípio para a seleção de vários modelos que explicassem a resposta de produtividade, no contexto da agricultura de precisão. Posteriormente, o critério de informação de Akaike foi utilizado para medir a parcimônia destes modelos.

O objetivo deste trabalho é apresentar os índices de desempenho mais utilizados e o critério de informação de Akaike, que contribuem para escolha do melhor modelo matemático para a representação de estudos espaciais. Neste trabalho será utilizado o software GEOEST (Vendrusculo et al., 2004), destinado à análise geoestatística e que disponibiliza tais índices. Com o intuito de exemplificar o ajuste e a escolha do melhor modelo do semivariograma, através dos índices de desempenho são utilizados os dados de precipitação anual média do Estado de São Paulo. Os dados são provenientes de mil e vinte e sete observações correspondentes as estações do DAEE, para o período de 1957 a 1997.

Índices de Desempenho e Critério de Informação de Akaike

Uma das maneiras de se encontrar a melhor estrutura de um modelo matemático para a representação de um sistema dinâmico pode se dar por meio da estimação de parâmetros para todas as possíveis estruturas e a conseqüente escolha baseado na comparação de alguns índices de desempenho.

No caso do semivariograma, os parâmetros estimados nos modelos matemáticos, são: C_o — Efeito pepita, C — patamar e a — alcance. Os modelos matemáticos mais usados no contexto agropecuário, que contemplam estudos de variabilidade das variáveis do solo e agroclimatológicos são os modelos esférico, exponencial e gaussiano, representados na Fig. 1.

[1] *Mestre em Engenharia Agrícola, Pesquisadora da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail:laurimar@cnptia.embrapa.br)*

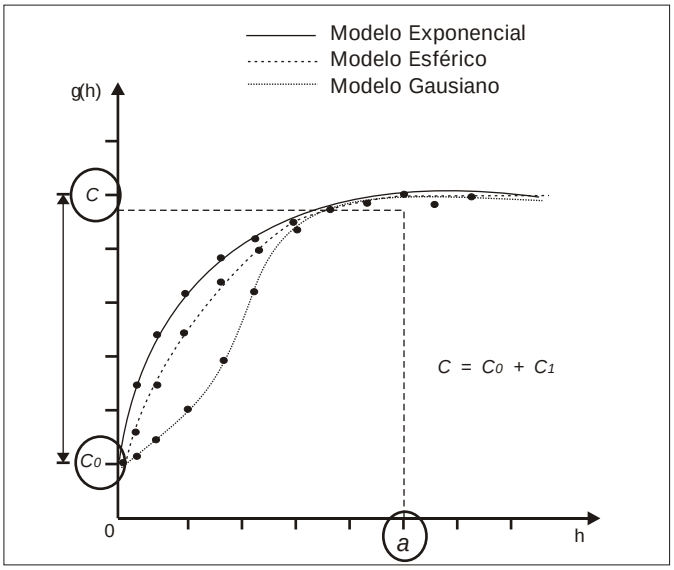


Fig. 1. Modelos teóricos de semivariograma.

O semivariograma pode ser definido como :

$$g^*(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(X_i) - Z(x_i + h)]^2 \quad (1)$$

onde N(h) é o número de pares de valores medidos Z(xi) e Z(xi + h), separados pela distância h se a variável for escalar (Carvalho et al., 2003).

O gráfico da Fig. 1 destaca em seus eixos (g(h) - semivariância e h - distância), na forma de círculos, os parâmetros do semivariograma. Esta ferramenta é uma função de h e portanto depende da magnitude e direção de h. O significado de cada parâmetro estrutural é descrito a seguir:

- Efeito Pepita (C_o) - quando o semivariograma tende a zero (g (h) = 0) o valor C_o é observado, revelando a descontinuidade do fenômeno para valores menores que a menor distância entre as amostras. Isaaks & Srivastava (1989) atribuem às condições adversas na época de medição ou a variabilidade de menor escala não percebida pelo processo de amostragem.
- Alcance (a): distância na qual as amostras se encontram correlacionadas espacialmente.
- Patamar (C): corresponde ao alcance (a) no gráfico do semivariograma. Considera-se que a partir deste ponto não exista mais dependência espacial entre as amostras. É aproximadamente igual a variância dos dados.
- Variância Estrutural (C1): é a diferença entre patamar e o efeito pepita.

Dentre os índices de desempenho freqüentemente utilizados para a tomada de decisão sobre o melhor modelo matemático cita-se:

1. Soma dos quadrados residual

$$SQR = \sum_{i=1}^{C(h)} (g(h)_{exp} - g(h)_{mod})^2 \quad (2)$$

onde g(h)exp corresponde a semivariância do semivariograma experimental (valores observados) e g(h)mod é a semivariância do modelo matemático (valores estimados). O modelo que apresenta o menor valor de RSS (RSS > 0) é assumido com sendo o melhor.

2. Coeficiente de determinação ou correlação múltipla (R^2) - o modelo que apresentar maior valor de R^2 ($0 < R^2 \leq 1$) é considerado o melhor. O valor de 1 para R^2 representa que o modelo teórico se adequou exatamente aos valores medidos no processo.

Pode-se calcular o coeficiente de determinação pela seguinte fórmula:

$$R^2 = \left(1 - \frac{\sum_{i=1}^{C(h)} (g(h)_{exp} - g(h)_{mod})^2}{\sum_{i=1}^{C(h)} (g(h)_{exp} - g(h)_{orig\ médio})^2}\right) \quad (3)$$

$g(h)_{exp}$ corresponde a semivariância do semivariograma experimental;

$g(h)_{mod}$ é a semivariância do modelo matemático;

$g(h)_{orig\ médio}$ é a semivariância média do semivariograma experimental.

3. Soma dos Quadrados dos Desvios Ponderados (SQDP) - neste caso escolhe-se o menor valor como o melhor. O cálculo deste índice é obtido pela seguinte fórmula:

$$SQDP = \sum_{i=1}^{C(h)} \left(\frac{P_i}{\sum_{i=1}^{C(h)} P_i} * (g_{(h)mod} - g_{(h)exp})^2\right) \quad (4)$$

onde:

P_i correspondem aos números de pares para cada classe de distância;

C(h) representa o número de classes de distâncias.

De acordo com Schaible et al. (1997), outros índices, não menos importantes, podem ser utilizados. São eles: Soma do quadrado total (Total sum of squares - TSS); Soma dos quadrados devido a regressão (Sum of square to regression - SS) e soma do quadrado dos resíduos normalizados (Normalized residual sum of square - NRSS).

Ressalta-se que o cálculo de R^2 e SQDP é realizado sobre todos os pontos do semivariograma experimental e do modelo. Na comparação entre os modelos procura-se valores de R^2 próximos à unidade e baixos valores de SQDP.

Estes critérios são largamente utilizados para a escolha do melhor modelo, porém não ponderam sobre o número de componentes usados para o modelo matemático estimado. Para tanto um compromisso satisfatório entre o bom ajuste e o princípio de parcimônia pode ser alcançado aplicando-se o chamado Critério de informação de Akaike (AIC), descrito em Akaike (1974) como um procedimento para identificação de modelo estatístico.

AIC = -2 log (Máxima Verossimilhança) + 2p (5)

Onde p é número de parâmetros do modelo independentemente ajustados.

Webster & McBratney (1989) utilizaram o AIC para escolha de modelos de variograma em propriedades de solos.

Disponibilidade Computacional dos Índices e Critério

São citados sumariamente a seguir dois sistemas computacionais que implementam os parâmetros sob estudo neste trabalho.

O Variogram Estimation and Spatial Prediction with Error - VESPER (Whelan et al., 2001), desenvolvido pelo Centro Australiano de Agricultura de Precisão da Universidade de Sidney e distribuído de forma shareware utiliza o Critério de Akaike e a soma do quadrado dos erros (Sum of square error) para a escolha do modelo de melhor ajuste dos dados.

Outra iniciativa baseada no trabalho de Vieira (1983) é o sistema GEOEST, desenvolvido em ambiente Delphi® seguindo o paradigma de orientação a objetos, para o ambiente Windows®. A Fig. 2 dá uma visão geral dos módulos do GEOEST e esclarece, por meio de um diagrama simplificado, em que fase são calculados os índices e o AIC.

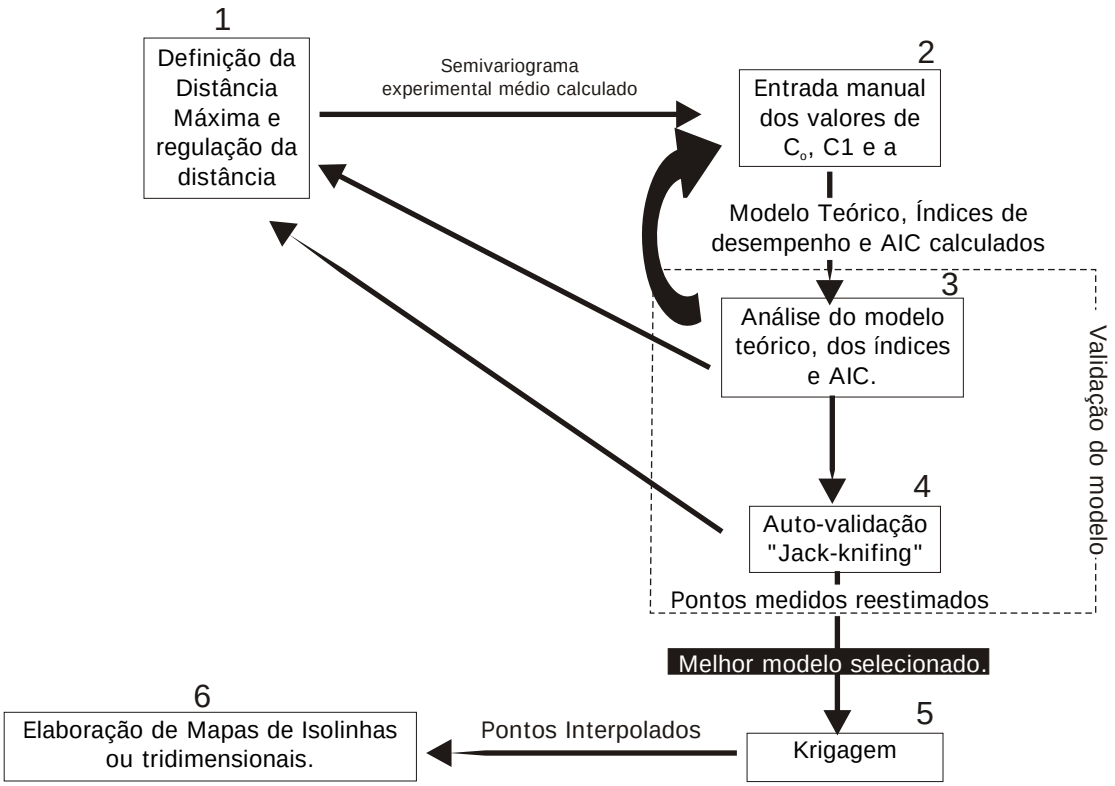


Fig. 2. Representação simplificada dos principais processos da análise geoestatística e seus respectivos subprodutos do GEOEST.

Estudo de Caso

Um exemplo, no software GEOEST, da utilização dos índices de desempenho e Critério de Akaike para dados de precipitação anual do estado de São Paulo é visualizado pela Fig.3.

O software destaca automaticamente o melhor valor de cada índice em cor vermelha, à medida que o usuário entra ou modifica os valores do alcance, efeito pepita e variância estrutural. A disponibilidade de tais informações auxilia a determinação dos parâmetros estruturantes do modelo a ser submetido à técnica de auto-validação *Jack-knifing*. Esta técnica reestima todos os pontos medidos em função dos pontos vizinhos e do modelo matemático escolhido. Devem ocorrer várias iterações (ajustes manuais) por parte do usuário optando pelo melhor modelo em função dos índices e da etapa de Jack-knifing, conforme destaca a Fig. 2. A atual versão do GEOEST não implementa a determinação automática dos valores de a, C1 e C_0. Esta funcionalidade pode ser programada, por exemplo, usando o método dos quadrados mínimos ponderados, descrito por Jian et al. (1996).

O software realiza o cálculo dos índices em modelos matemáticos de estrutura simples, ou seja, não aninhados. Nos modelos aninhados, como por exemplo duplo esférico, devem haver tratamento específico no que se refere ao AIC (Webster & McBratney, 1989).

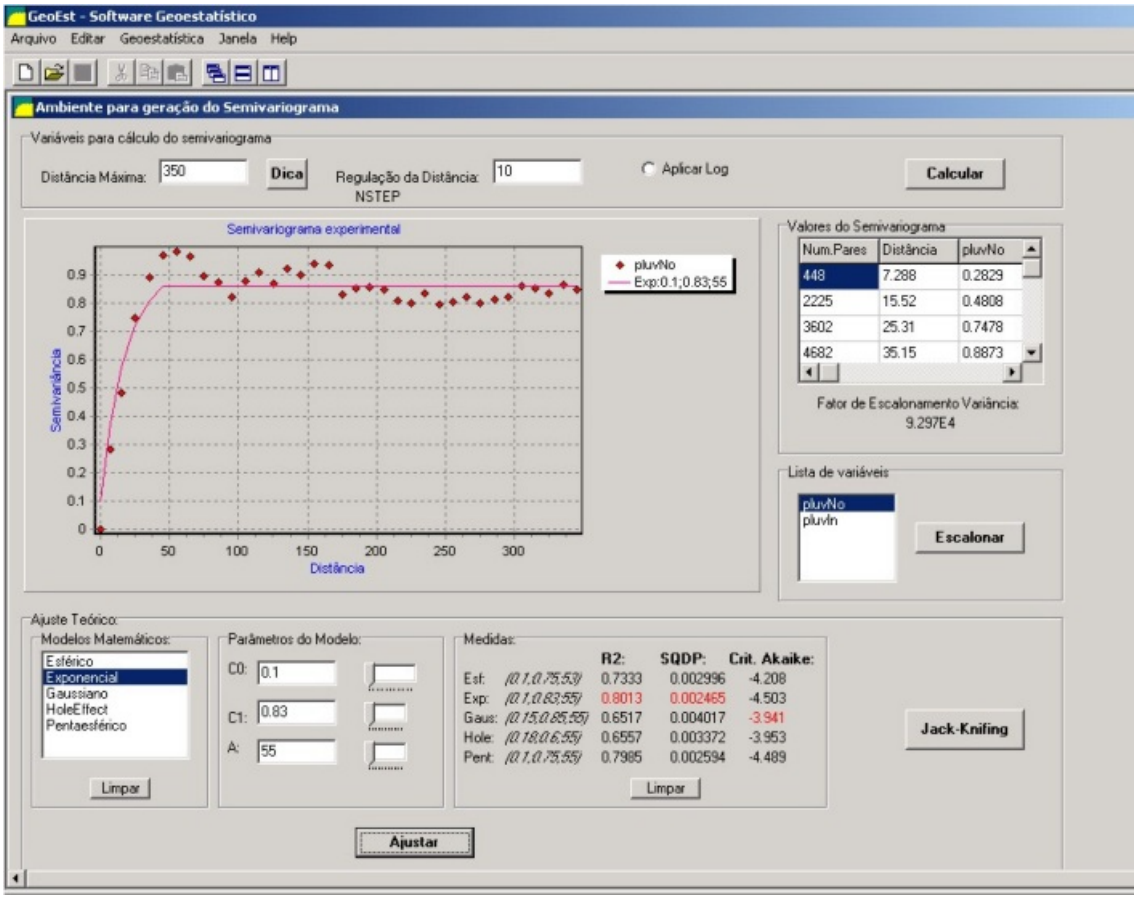


Fig. 3. Ambiente para modelagem do semivariograma destacando os índices de desempenho (R^2 e SQDP) e Critério de Akaike (AIC) para dados de precipitação no estado de São Paulo.

A situação ideal seria que os índices de desempenho e AIC convergissem para um único modelo dos cinco apresentados pela Fig. 3. Ou seja o melhor modelo seria aquele que apresentasse, concomitantemente, o maior valor de R^2, menor valor de SQDP e menor valor de AIC. Na Fig. 3, o menor valor de AIC corresponde ao modelo Gaussiano (-3.941) e os melhores valores de R^2 e SQDP coincidem com o modelo exponencial.

Na situação mostrada pela Fig. 3 mais iterações dos parâmetros do modelo seriam necessárias ou pelo conhecimento da variabilidade do atributo, o usuário poderia optar pela escolha do modelo matemático que atendesse a maioria dos índices.

Conclusões

- Uso de índices de desempenho e critério de informação de Akaike constituem-se em importante instrumento utilizado para a escolha do melhor modelo para ajuste do semivariograma, diminuindo a subjetividade que permeia este processo. Esse fato influencia diretamente a melhoria da confiança dos valores interpolados, pois o modelo escolhido representa com mais fidelidade a variabilidade do atributo estudado; visto isso, recomenda-se fortemente aos usuários da técnica geoestatística testar os dados de seu interesse em sistemas que implementem algum ou todos estes índices e critério; e
- para aplicações que permitam a modelagem do semivariograma de forma manual e onde são necessárias grande número de interações a disponibilidade destes índices e do critério no Software GEOEST, e o destaque automático dos melhores valores agiliza o processo de modelagem do semivariograma, etapa fundamental na análise de variabilidade geoestatística.

Referências Bibliográficas

AKAIKE, H. A new look at statistical model identification. IEEE Trans. on Automatic Control, v. 19, n. 6, p. 716-723, 1974.

CARVALHO, J. R. P. de; QUEIROZ, E. F. de; VIEIRA, S. R. Uso da geoestatística multivariada com incorporação de altitude na interpolação espacial da precipitação. In: CONGRESSO BRASILEIRO DE CIÊNCIA DO SOLO, 29., 2003, Ribeirão Preto. [Anais...]. Ribeirão Preto: Sociedade Brasileira de Ciência do Solo, 2003. CD-ROM.

ISAAKS, E. H.; SRIVASTAVA, R. M. An introduction to applied geostatistics. New York: Oxford University Press, 1989. 561 p.

JIAN, X.; OLEA, R. A.; YU, Y. Semivariogram modeling by weighted least squares. Computers & Geosciences, v. 22, n. 4, p. 387-397, 1996.

LARK, R. M. Some tools for parsimonious modelling and interpretation of within-field variation of soil and crop system. Soil & Tillage Research, v. 58, n. 3-4, p. 99-111, 2001.

SCHAIBLE, B.; XIE, H.; LEE, Y. C. Fuzzy logic models for ranking process effects. IEEE Transactions on Fuzzy Systems, v. 5, n. 4, p. 545-556, 1997.

VENDRUSCULO, L. G.; MAGALHÃES, P. S. G.; VIEIRA, S. R.; CARVALHO, J. R. P. de. Computational system for geostatistical analyses. Scientia Agricola, Piracicaba, v. 61, n. 1, p. 100-107, 2004.

VIEIRA, S. R. Geoestatística em estudos de variabilidade espacial do solo. In: NOVAIS, R. F. de; ALVAREZ V. V. H; SCHAEFER, C. E. G. R. (Ed.). Tópicos em ciência do solo. Viçosa: Sociedade Brasileira de Ciência do Solo, 2000. v. 1, p. 1-54.

VIEIRA, S. R.; HATFIELD, T. L.; NIELSEN, D. R.; BIGGAR, J. W. Geostatistical theory and application to variability of some agronomical properties. Hilgardia, Berkeley, v. 51, n. 3, p. 1-75, 1983.

WEBSTER, R.; MCBRATNEY, A. B. On the Akaike information criterion for choosing models for variograms of soil properties. Journal of Soil Science, v. 40, n. 3, p. 494-496, 1989.

WHELAN, B. M; MCBRATNEY, A. B.; MINASNY, B. VESPER - spatial prediction software for precision agriculture. In: GRENIER, G.; BLACKMORE, S. (Ed.). Proceedings of the 3rd. European Conference on Precision Agriculture. Montpellier: Agro-Montpellier, 2001. p. 139-144.

Comunicado Técnico, 58

Ministério da Agricultura, Pecuária e Abastecimento

Governo Federal

Embrapa Informática Agropecuária
Área de Comunicação e Negócios (ACN)
Endereço: Caixa Postal 6041 - Barão Geraldo
13083-970 - Campinas, SP
Fone: (19) 3789-5743
Fax: (19) 3289-9594
e-mail: sac@cnptia.embrapa.com.br

1ª edição on-line - 2003

Expediente

Supervisor editorial: *Ivanilde Dispato*
Normalização bibliográfica: *Marcia Izabel Fugisawa Souza*
Editoração eletrônica: *Área de Comunicação e Negócios*